Assessing Classifier Fairness with Collider Bias

Zhenlong Xu[1], Ziqi Xu[1], Jixue Liu[1], Debo Cheng[1], Jiuyong Li[1(✉)], Lin Liu[1], and Ke Wang[2]

[1] University of South Australia, Adelaide, Australia
{Zhenlong.Xu,Ziqi.Xu,Debo.Cheng}@mymail.unisa.edu.au,
{Jixue.Liu,Jiuyong.Li,Lin.Liu}@unisa.edu.au
[2] Simon Fraser University, Burnaby, Canada
wangk@sfu.ca

Abstract. The increasing application of machine learning techniques in everyday decision-making processes has brought concerns about the fairness of algorithmic decision-making. This paper concerns the problem of collider bias which produces spurious associations in fairness assessment and develops theorems to guide fairness assessment avoiding the collider bias. We consider a real-world application of auditing a trained classifier by an audit agency. We propose an unbiased assessment algorithm by utilising the developed theorems to reduce collider biases in the assessment. Experiments and simulations show the proposed algorithm reduces collider biases significantly in the assessment and is promising in auditing trained classifiers.

Keywords: Fairness · Collider bias · Causal inference

1 Introduction

There are increasing concerns over the fairness of decision making algorithms with the wide use of machine learning in various applications, such as job hiring, credit scoring and home loan since discrimination can be inadvertently introduced into machine learning models. To prevent unfairness in a model from spreading in society, audit techniques are needed for the independent authority to audit machine learning models. Figure 1 shows an audit process. An audit agency accesses a model of a company and has its own audit cases for assessing the fairness of the model. The audit agency does

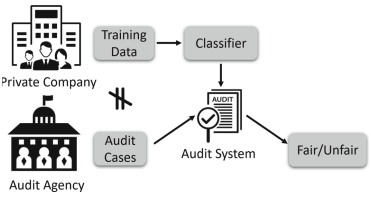

Fig. 1. The process of audit.

Supported by Australian Research Council (DP200101210), Natural Sciences and Engineering Research Council of Canada and Postgraduate Research Scholarship of University of South Australia.

Z. Xu and Z. Xu—contributed equally to this paper.

J. Gama et al. (Eds.): PAKDD 2022, LNAI 13281, pp. 262–276, 2022.
https://doi.org/10.1007/978-3-031-05936-0_21

not have access to the training data set but has the regulatory policy. In this paper, we use a causal graph to represent the regulatory policy. The company may use additional variables that are not specified in the regulatory policy to build its models to improve prediction accuracy.

Situation test has been used in the U.S. to detect discrimination in recruitment [2], which is a controlled experiment approach for analysing employers' decisions on job applicants' characteristics, as illustrated with the following examples. Pairs of research assistants are sent to apply for the same job, and each pair of the pretended applicants have the same qualifications and experience related to the job but have different values for their protected variable, such as male/female or young/old. Discrimination is detected if the favourable decisions are unequal between groups with different protected values.

The above described situation test can be simulated in an audit process, and we call it Naive Situation Test (NST) in this paper. We feed two inputs representing two individuals whose variable values are identical except their protected values to a machine learning model. If the model provides different decisions, NST will detect the model as discriminatory.

NST may produce an incorrect detection. We use the following example to show this. Consider a classifier $f()$ used by a company to determine employees' salaries as $salary = f(race, education, suburb)$. Some predicted outcomes by the model are shown in Table 1. Based on NST, the black people are not discriminated against since with the same education and suburb, both white and black people are predicted to have the same salary. However, Suburb is an irrelevant variable for determining the Salary. Without considering the Suburb, with the same level of Education, 2/3 white people receive a salary higher than 50K while only 1/3 black people receive a salary of 50K or higher. Hence, black people are discriminated against by the model.

Table 1. An example of incorrect detection by NST on a classifier.

Race	Edu	Sub	Predicted.Sal
white	high	A	>50k
black	high	A	>50k
white	high	A	>50k
black	high	B	≤50k
black	high	B	≤50k
white	high	B	≤50k

NST ⇒ "fair"

f(**white**, high, A)=f(**black**, high, A)
f(**white**, high, B)=f(**black**, high, B)

The incorrect detection by NST is caused by collider bias. We use a causal graph, formally defined in Sect. 2, to explain the collider bias. Causal relationships of variables in the above example are shown using the causal graph in Fig. 2 where a directed edge represents a causal relationship. The suburb is a collider since two edges "collide" at it. Conditioning on a collider, an association is formed between the two variables but it is spurious [5]. In the example, the spurious association cancels the association due to the causal relationship between Race and Salary and hides the true discrimination. Collider bias is related to the selection bias [10]. In a classifier, conditioning on a variable is equivalent to selecting sub-populations using the values of the variable. If the

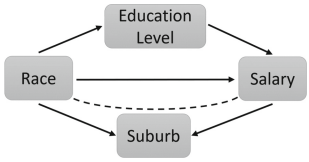

Fig. 2. The causal graph for the above example.

variable is a collider, a selection bias in each sub-population is resulted. We call this bias collider bias in this paper.

There is no existing alternative method to NST to audit classifiers. Most methods, to be reviewed in the related work, need to access the training data set and suffer from collider bias. A causal-based situation test (CST) [25] does not suffer from collider bias, but needs to access the training data set too. The audit cases used by an audit agency are only a small number of individual cases which do not represent the population. Collecting a representative sample of the population needs a significant resource. Therefore, a data-based audit method is not applicable. We make the following contributions in this paper.

- We study collider bias in fairness assessment and present theorems to avoid collider bias. Our theoretical results give a principled guidance on which variables can be used for fairness assessment and also for building fair classifiers.
- We investigate the problem of auditing machine learning models and propose an Unbiased Situation Test (UST) algorithm for auditing without accessing training data or an unbiased sample of the population. Experiments show that UST can effectively reduce collider bias.

2 Background

We present the necessary background of causal inference. We use upper case letters to represent variables and bold-faced upper case letters to denote sets of variables. The values of variables are represented using lower case letters.

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph, where $\mathbf{V} = \{V_1, \ldots, V_p\}$ is the set of nodes and $\mathbf{E}$ is the set of edges between the nodes, i.e. $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. A path π is a sequence of distinct nodes such that every pair of successive nodes are adjacent in $\mathcal{G}$. A path π is a directed path if all edges along the path are directed edges. A path between (V_i, V_j) is a backdoor path with respect to V_i if it has an arrow into V_i. Given a path π, V_k is a collider node on π if there are two edges incident like $V_i \rightarrow V_k \leftarrow V_j$. In $\mathcal{G}$, if there exists $V_i \rightarrow V_j$, V_i is a parent of V_j and we use $Pa(V_j)$ to denote the set of all parents of V_j. In a directed path π, V_i is an ancestor of V_j and V_j is a descendant of V_i if all arrows point to V_j.

A DAG (Directed Acyclic Graph) is a directed graph without directed cycles. With the following two assumptions, a DAG links to a distribution.

Definition 1 (Markov condition [17]). *Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and $P(\mathbf{V})$, the joint probability distribution of $\mathbf{V}$, $\mathcal{G}$ satisfies the Markov condition if for $\forall V_i \in \mathbf{V}$, V_i is probabilistically independent of all non-descendants of V_i, given the parents of V_i.*

When the Markov condition holds, $P(\mathbf{V})$ can be factorised into: $P(\mathbf{V}) = \prod_i P(V_i \mid Pa(V_i))$.

Definition 2 (Faithfulness [20]). *A DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is faithful to $P(\mathbf{V})$ iff every independence presenting in $P(\mathbf{V})$ is entailed by $\mathcal{G}$ which fulfills the Markov condition. A distribution $P(\mathbf{V})$ is faithful to a DAG $\mathcal{G}$ iff there exists DAG $\mathcal{G}$ which is faithful to $P(\mathbf{V})$.*

With the above two assumptions, we can read the independencies between variables in $P(V)$ from a DAG using the Definition 8 in Appendix A. To conduct causal inference with DAGs, we make the following assumptions.

Definition 3 (Causal sufficiency [20]). *A data set satisfies causal sufficiency if for every pair of variables (V_i, V_j) in $\mathbf{V}$, all their common causes are also in $\mathbf{V}$.*

With a DAG, if we interpret a node's parent as its direct cause, the DAG is known as a causal DAG. We can learn a causal DAG from data when the assumptions of causal sufficiency, faithfulness and Markov condition are satisfied.

An intervention, which forces a variable to take a value, can be represented by a *do* operator. For example, $do(X = 1)$ means X is intervened to take value 1. $P(y \mid do(X = 1))$ is an interventional probability. Let us understand *do* in an ideal experiment.

Definition 4 (Direct effect [17]). *The direct effect of X on Y is $P(y \mid do(X = x), do(\mathbf{V}_{\setminus XY} = \mathbf{v}))$ where $\mathbf{V}_{\setminus XY}$ means all other variables except X and Y.*

In order to study the relationship between X on Y, all other variable are controlled in the ideal experiment. To infer interventional probabilities (by reducing them to normal conditional probabilities) with a causal DAG, the rules of *do*-calculus [17] are necessary. Detailed description of these rules are available in Appendix A, and we used these rules to proof our theorems.

3 Problem Definition

A classifier (prediction model) has been built by a company/organisation from a training data set which contains a binary protected variable A, a binary decision outcome Y, and a set of relevant variables of Y, $\mathbf{X}$, since variables independent of Y are not used for predicting Y. An agency wants to audit the model using some cases. We make the following assumptions about the audit.

Assumption 1 *1. The regulatory policy has specified the causal relationships among the factors and Y, and uses a causal DAG to indicate. The factors are ancestral variables of Y including all direct causes of Y.*
2. The audit agency has no access to the model training data or an unbiased sample of the population. The agency however has access to the distributional statistics from some sources, such as government census data.
3. The company or organisation has used all the legitimate factors to comply with the regulatory policy. However, some other variables are also used by the model to enhance the prediction performance.

In the theorem development, we assume that there is a DAG that is consistent with the regulatory policy. In the algorithm, we do not need the complete DAG, but ancestral variables of Y and colliders in the descendant nodes of Y.

We first define the criterion for auditing. We use Controlled Direct Effect (CDE) [18] to measure fairness. CDE is extended Definition 4 to simulate an ideal experiment. The alternative definitions are path specific causal effect [4,21] and counterfactual fairness [12], we will discuss why the alternatives have not been used after Definition 5.

The protected variables in this paper include redline variables, which are the descendants of protected variables. The redline variables are recognised as a proxy of protected variables and may cause some discrimination [11]. Some companies or organisations build the models under the concept of fairness through awareness [7], which means the classifier functions may not use the protected variables as input. In this case, the redline variables will be considered as the protected variables.

Definition 5. (Fairness score). *Given a causal DAG $\mathcal{G}$ representing the regulatory policy, A, $\mathbf{X}$, and Y as described above. The fairness score is of an individual (or a subgroup) $\mathbf{X} = \mathbf{x_i}$ is defined by Controlled Direct Effect, $CDE(\mathbf{x_i}) = P(y \mid do(A = 1), do(\mathbf{X} = \mathbf{x}_i)) - P(y \mid do(A = 0), do(\mathbf{X} = \mathbf{x}_i))$, where y denotes $Y = 1$.*

The rationale of the above definition is that we conduct a controlled experiment by intervening the protected variable, and controlling all other variables to $\mathbf{x}_i$. The decision for $\mathbf{x}_i$ is fair if the intervention does not change the outcome.

Unlike previous works [7–9,15], our definition of fairness score is based on the CDE which uses intervention. Thus the spurious association between A and Y caused by conditioning on colliders will be avoided. We do not use counterfactual fairness [12] in our fairness definition since it needs stronger assumptions and poses a practical challenge. To estimate counterfactual outcomes, there is a need for knowing the full causal model and latent background knowledge. Both are not available in our problem setting. Some other definitions [4,21] make use of path specific causal effect. Their solutions also need counterfactual reasoning and they do not fit our problem setting.

Definition 6. (Problem definition). *Given $\mathcal{G}$, A and $\mathbf{X}$ as described above, and classier $\hat{Y} = f(A, \mathbf{X})$. The audit is to determine if a prediction on an individual ($\mathbf{X} = \mathbf{x_i}$) is fair, i.e. $|CDE(\mathbf{x_i})| < \tau$ where τ is a threshold determined by the regulatory policy and Y in $CDE(\mathbf{x_i})$ is replaced by $\hat{Y}$.*

4 Estimating CDE

For the sake of fairness audit, the protected variable A is assumed to be a parent node of Y so we can use CDE for the audit. The results in this section are true in general, not just for auditing classifiers. Due to page limitations, all the proofs of theorems will be presented in Appendix B.

Theorem 1. *DAG $\mathcal{G}$ contains variables A and Y, and variable set $\mathbf{X}$ where $(A \cup Y) \cap \mathbf{X} = \emptyset$. The causal sufficiency is satisfied. $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Pa'(Y) = \mathbf{pa})$ where $Pa'(Y)$ is the set of all parents of Y in $\mathcal{G}$ excluding A.*

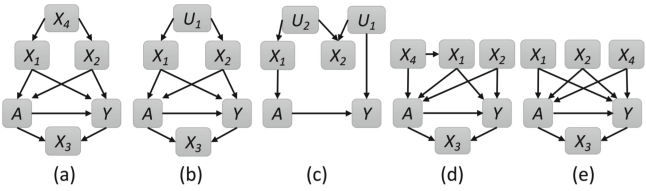

Fig. 3. DAGs for the examples of Theorem/Corollary. X_1 to X_6 are observed variables, and U_1 and U_2 are unobserved variables.

Theorem 1 removes the descendant nodes of Y from the conditioning set in the conditional probabilities for CDE estimation, and this removes possible collider bias. Furthermore, it gives a succinct set of variables for estimating CDE.

For example, in Fig. 3(a), $P(y \mid do(a), do(x_1, x_2, x_3, x_4)) = P(y \mid a, x_1, x_2)$ based on Theorem 1, where we use x_i for $X_i = x_i$. The CDE is determined by conditional probabilities on A, X_1 and X_2. Since X_3 is not used in the conditioning set, there will be no collider bias. Theorem 1 is based on the causal sufficiency assumption, which assumes that there are no unobserved common causes in the data set. In real-world applications, unobserved variables are unavoidable. When there are unobserved variables, how do we estimate CDE? The following corollary will show that they do not invalidate the result of Theorem 1.

Corollary 1. *Let $Ca(Y)$ include all the direct causes and only direct causes of Y except A. $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Ca(Y) = \mathbf{ca})$.*

Corollary 1 indicates that discrimination detection is sound when the audit agency knows all the direct causes of Y and uses them as the conditioning set when calculating CDE. For example, in Fig. 3(b), $P(y \mid do(a), do(x_1, x_2, x_3, U_1)) = P(y \mid a, x_1, x_2)$ based on Corollary 1. Unobserved ancestral variables of Y are blocked off from Y by X_1 and X_2, and they do not affect the probability of Y. The unobservable variables can be in the descendant nodes of Y too, but they do not affect the CDE estimation since they will not be used anyway.

We will further explain why direct causes are necessary for Corollary 1. Let Fig. 3(c) be a true DAG with two unobserved variables U_1 and U_2. X_2 is not a direct cause of Y. Since U_1 and U_2 are unobserved, X_2 is perceived as a parent of Y in the observed data. If X_2 is used to estimate CDE, the estimation will be biased since the back door path (Y, U_1, U_2, X_1, A) is opened when X_2 is conditioned on. In this case, X_1 is necessary to block the path. When both X_1 and X_2 are included, the CDE estimation is unbiased. Sometimes, we need redundancy to prevent such a biased estimation.

Both Theorem 1 and Corollary 1 give a succinct conditioning set for CDE estimation. In fact, a superset of the direct causes works as long as the superset does not contain descendant nodes of Y. In a DAG, ancestral nodes represent the direct causes and indirect causes of Y.

Corollary 2. *Let* $\mathbf{B}$ *include all the direct causes and some (or all) indirect causes of* Y*. We have* $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, \mathbf{B} = \mathbf{b})$.

Corollary 2 allows some redundancy in the conditioning set comparing to Corollary 1. In practice, the redundancy gives a flexibility for users to determine the direct causes of Y. Sometimes, a direct cause and an indirect cause are difficult to distinguish, and Corollary 2 indicates that including both does not bias the CDE estimation. For example, in Fig. 3(d), $P(y \mid do(a), do(x_1, x_2, x_3, x_4)) = P(y \mid a, x_1, x_2) = P(y \mid a, x_1, x_2, x_4)$ based on Corollary 2 if X_1 and X_2 are all direct causes of Y. Let us assume that Fig. 3(d) is the true DAG, but a government agency has a DAG as Fig. 3(e) since they do not know which one of X_1 and X_4 is the direct cause of Y. A CDE estimation based on the imprecise DAG in Fig. 3(e), i.e. $P(y \mid do(a), do(x_1, x_2, x_3, x_4)) = P(y \mid a, x_1, x_2, x_4)$ is also unbiased.

5 Implementing Unbiased Situation Test

We summarise the discussion and propose the following unbiased situation test.

Definition 7 (Unbiased Situation Test (UST)). *UST exams whether a classifier* $\hat{Y} = f()$ *is fair for a given case* $\mathbf{x_i}$ *by calculating* $CDE(\mathbf{x_i}) = P(\hat{Y} = 1 \mid A = 1, \mathbf{B} = \mathbf{b_i}) - P(\hat{Y} = 1 \mid A = 0, \mathbf{B} = \mathbf{b_i})$*, where* $\mathbf{B}$ *is the set of direct causes and some (or all) indirect causes of* Y*. The test case* $\mathbf{x}_i$ *is discriminated if* $|CDE(\mathbf{x_i})| \geq \tau$*, where* τ *is a threshold specified by the regulatory policy.*

All variables in the problem except (A, Y) can be categorized into two types: $\mathbf{B}$ and $\mathbf{C}$. $\mathbf{B}$ is the set of ancestral nodes of Y which can be identified by the regulatory policy, and $\mathbf{C}$ includes others. Note that irrelevant variables which are independent of Y are not in $\mathbf{X}$.

To conduct UST as in Definition 7, one problem is that an audit agency cannot obtain the conditional probability $P(\hat{Y} = 1 \mid A = a, \mathbf{B} = \mathbf{b}_i)$ directly since it does not access the training data set or a unbiased sample of the population.

Algorithm 1. Unbiased Situation Test (UST)

Input: Classifier $f()$, $\mathbf{X} = \mathbf{B} \cup \mathbf{C}$ as defined in the text. $P(\mathbf{C} = \mathbf{c_i})$. Test cases D_{Test}. The threshold τ.
Output: L, a list of discriminated cases in D_{Test}.

1: **for** each $r_i \in D_{Test}$ **do**
2: Let r_i' be the record by flipping the value of A in r_i
3: Let $P(\hat{Y} = 1 \mid r_i) = f(r_i)$ and $P(\hat{Y} = 1 \mid r_i') = f(r_i')$
4: Obtain $P(\hat{Y} = 1 \mid A = A(r_i), \mathbf{B} = B(r_i))$ and $P(\hat{Y} = 1 \mid A = A(r_i'), \mathbf{B} = B(r_i'))$ by Equation 1 where $A()$ and $B()$ return values of A and $\mathbf{B}$ in the records respectively
5: Conduct situation test by Definition 7 and update L
6: **end for**
7: Return L

Instead, it can have $P(\hat{Y} = 1 \mid A = a, \mathbf{B} = \mathbf{b_i}, \mathbf{C} = \mathbf{c_i})$ from the classifier $f()$. Therefore, the following marginalisation is used:

$$P(\hat{y} \mid A = a, \mathbf{B} = \mathbf{b_i}) = \sum_{\mathbf{c_i} \in \mathbf{C}} P(\hat{y} \mid a, \mathbf{b_i}, \mathbf{C} = \mathbf{c_i}) P(\mathbf{c_i}) \tag{1}$$

where $\hat{y}$ denotes $\hat{Y} = 1$, and probability $P(\mathbf{c_i})$ can be obtained from some sources, such as government census data. The algorithm for UST is presented in Algorithm 1. The complexity for UST algorithm is $O(n)$, where n is the size of D_{Test}, i.e. linear to the number of test records.

6 Experiments

In this section, we first demonstrate UST algorithm can correct the spurious associations generated by collider. Then, we simulate the audit process by using real-world data set. We only compare UST with NST in population-level sampling since other situation test methods, such as, CST [25], k-NN based situation test [15] need to access training data set which is unavailable in our problem. We also demonstrate that data-based audit method fails in unrepresentative sampling but UST works. Finally, we apply the UST algorithm to compare fairness for different models and guide the audit agency to choose the model. The experimental settings and details can be found in the full version [22].

6.1 Correcting Collider Biases

We construct synthetic data sets including a collider as discussed in the full version [22]. UST has significantly reduced biases in CDE estimation. Bias is used to measure the error between an estimated CDE and the true CDE. Biases of including and not including a collider are shown in Table 2. The later is the UST method which corrects biases of a collider in data by directly using Corollary 2.

Table 2. UST has significantly reduced biases caused by a collider.

Trials	Bias (with collider)	Bias (UST)
1	0.143 ± 0.011	0.072 ± 0.004
2	0.154 ± 0.013	0.074 ± 0.004
3	0.149 ± 0.012	0.066 ± 0.004
4	0.149 ± 0.014	0.067 ± 0.004
5	0.152 ± 0.012	0.069 ± 0.004

Table 3. Suburb variable (collier) improves the accuracy of classification models.

	Acc. w/ Sub	Acc. w/o Sub
DT	89.66%	81.01%
SVM	89.60%	80.93%
RF	89.81%	80.86%
NN	89.64%	80.85%

6.2 Simulating an Audit Process Using Adult Data Set

The Adult data set from UCI Machine Learning Repository [1] is used to simulate audit process as shown in Fig. 4. We use the Adult data set as the population for generating the ground truths. A company has a sample (50%) as the private data to build a model. The red dashed line represents the information that the audit agency has access to. The ground truths are generated from the population and all causes of Y.

Race is the protected variable and Salary is the outcome. Other variables are Education_level, Marriage_statues, Work_hour, Work_class, and they all determine the salary. We simulate a Suburb variable as a collider. The accuracy of a classifier is significantly higher when the Suburb is used than not as shown in Table 3. The accuracy improvement is due to the spurious associations.

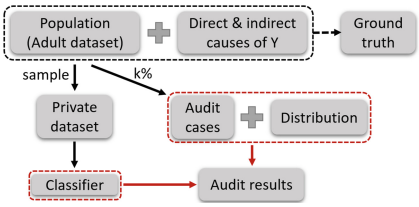

Fig. 4. A simulation of audit process using Adult data set

6.3 Comparing the Audit Performance of NST and UST

We apply UST to audit a few well-known classifiers built from sample data set. NST (introduced in the introduction) is used for the comparison since it is the only method for assessing the fairness of a classifier without accessing the training data set. We use precision and recall for the comparison. The ground truth for each audit case is calculated by using the population and the causes of Y. Audit cases are k% records randomly selected from the population. For each k, we resample audit cases 10 times and report the average precision and recall.

UST outperforms NST in both precision and recall as shown in Table 4. With the increasing number of audit cases, the deviations of both methods decrease. From the gaps between the precision and recall of NST and UST, we see that the collider bias deteriorates the detection performance of NST significantly.

Table 4. The audit performance comparison of NST and UST. The higher values are highlighted. The standard error is shown in brackets.

		k = 0.1%		k = 0.5%		k = 1%	
		NST	UST	NST	UST	NST	UST
DT	Recall	59.6%(0.96)	**79.8%(0.11)**	56.7%(0.27)	**73.3%(0.05)**	56.3%(0.05)	**71.7%(0.10)**
	Precision	84.4%(0.32)	**98.1%(0.16)**	78.9%(0.06)	**99.1%(0.01)**	80.3%(0.02)	**98.8%(0.01)**
SVM	Recall	77.9%(1.18)	**87.8%(0.26)**	75.7%(0.17)	**89.9%(0.03)**	74.1%(0.06)	**89.1%(0.02)**
	Precision	68.1%(0.73)	**83.4%(0.14)**	65.1%(0.11)	**79.9%(0.06)**	64.6%(0.13)	**81.0%(0.05)**
RF	Recall	56.1%(1.91)	**73.8%(0.32)**	58.2%(0.23)	**66.8%(0.03)**	57.7%(0.09)	**65.2%(0.14)**
	Precision	88.6%(0.17)	**96.4%(0.25)**	86.8%(0.02)	**97.8%(0.01)**	86.9%(0.04)	**98.4%(0.01)**
NN	Recall	65.9%(1.17)	**74.9%(0.14)**	67.6%(0.16)	**71.5%(0.10)**	67.9%(0.06)	**69.2%(0.08)**
	Precision	85.9%(0.24)	**96.7%(0.21)**	81.7%(0.04)	**97.1%(0.01)**	82.8%(0.02)	**97.2%(0.01)**

6.4 Data Based Audit May Be Biased

Removing the collider from the data can be used as an alternative method to UST. However, a data-based audit (DBA) relies on the representativeness of the audit cases for the population. The representativeness is difficult to be ensured because individuals who receive unfair treatments likely require the authority to audit their results. An audit agency does not have a resource to collect a representative sample for auditing. We simulate the unrepresentative audit cases by over (under) sampling discriminatory cases in the population. In the Adult data set, about 15% of the individuals are discriminatory and this ratio is the baseline. We vary discriminatory ratios of 1% data set.

The performance of DBA deteriorates significantly when a discriminatory ratio deviates from the baseline as shown in Table 5. Note that all discrimination detection methods based on data have the same limitation. In contrast, UST maintains similar performance.

Table 5. The audit performance comparison of DBA and UST with discriminatory sample. The higher values are highlighted. The standard error is shown in brackets.

		Discriminatory Ratio=0%		Discriminatory Ratio=10%		Discriminatory Ratio=20%	
		DBA	UST	DBA	UST	DBA	UST
DT	Recall	60.1%(1.18)	**72.2%(0.06)**	57.8%(1.72)	**71.7%(0.07)**	60.9%(0.91)	**71.8%(0.10)**
	Precision	84.6%(0.23)	**97.5%(0.01)**	78.2%(0.35)	**96.2%(0.01)**	70.9%(0.11)	**95.1%(0.01)**
SVM	Recall	39.8%(1.64)	**89.0%(0.01)**	40.0%(1.51)	**89.5%(0.01)**	35.6%(2.57)	**89.8%(0.01)**
	Precision	77.8%(0.19)	**82.5%(0.03)**	71.2%(0.31)	**76.0%(0.05)**	56.4%(1.02)	**67.7%(0.06)**
RF	Recall	39.9%(0.18)	**65.6%(0.06)**	40.4%(0.14)	**65.2%(0.06)**	39.5%(1.13)	**65.2%(0.05)**
	Precision	78.9%(0.14)	**97.3%(0.02)**	72.6%(0.16)	**96.4%(0.02)**	61.2%(0.26)	**94.2%(0.05)**
NN	Recall	46.1%(2.02)	**69.8%(0.07)**	45.7%(2.11)	**69.8%(0.06)**	39.4%(2.26)	**69.4%(0.09)**
	Precision	80.6%(0.11)	**95.0%(0.03)**	73.6%(0.29)	**93.0%(0.03)**	60.0%(0.51)	**90.1%(0.07)**

6.5 Rank Models Based on Fairness

We show that UST can be used for comparing the fairness of different models. We first discuss the metrics for the comparison. After discrimination detection on a model using audit cases with ground truths, we obtain True Positive (TF), False Positive (FP), True Negative (TN), and False Negative (FN). FN indicates the cases that are unfair but are corrected to be fair by the model. They are favourable for fair predictions, and we use correction rate, $CR = \frac{FN}{TP+FN}$, to represent the proportion of true unfair cases being corrected by a model. In contrast, FP represents that the cases that are fair become unfair after model predictions. These cases are called reversed discrimination and are unfavourite for predictions. We use the reversion rate, $RR = \frac{FP}{FP+TN}$, to represent the proportion of fair cases being reversed by a model. We wish the revision rate is as small as possible.

The CR and RR of four classifiers are shown in Table 6. Random Forest is the fairest model based on the two measures. Random Forest has corrected 35.24% of unfair cases, and only reversed 1.07% of fair cases to unfair. Note that, their prediction accuracies are very similar, but their CR and RR are different. This assessment shows that some errors made by a model are better than others in terms of the fairness.

Table 6. The audit results of various models.

	CR(↑)	RR(↓)
DT	29.09% ± 0.079	1.69% ± 0.001
SVM	10.05% ± 0.024	39.37% ± 0.049
RF	35.24% ± 0.147	1.07% ± 0.001
NN	31.48% ± 0.064	2.76% ± 0.001

7 Related Works

The work belongs to discrimination detection. Detection methods are divided into the group, and individual-based. Another division of the methods is association or causal based.

At the group level, a number of metrics have been defined to detect discrimination. Demographic parity, a well-known fairness measurement, is defined by [7]. Other measurements including equalised odds [9], predictive rate parity [23]. However, these group-based fairness does not necessarily mean individual fairness. Many algorithms focus on detecting discrimination at the individual level. Authors in [19] use existing inequality indices from economics to measure individual level fairness. Speicher et al. [14] propose an individual level discrimination detector, which is used to prioritise data samples and aims to improve the subgroup fairness measure of disparate impact.

Under the causal framework, Li et al. [13] use the (conditional) average causal effect to quantify fairness for (sub)group level discrimination detection. Counterfactual fairness [12] is an attractive definition of individual level fairness measurements by causality. It means that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world (when a value of a protected variable is changed). However, it needs strong assumptions. Zhang et al. [26] use nature direct effect and nature indirect effect to quantify fairness. The path-specific causal effect [4,21] have been used to quantify fairness when the regulatory policy recognises some causal paths involving a protected variable fair. Nature direct (indirect) effect and path-specific causal effect all need counterfactual reasoning and are difficult to implement in practice since the strong assumptions are related to counterfactual reasoning.

Situation test related work has been discussed in the introduction. The abovementioned related work only introduces some main influential contributions. For more related work, please refer to the literature review [3,6,16,24].

8 Conclusions

In this paper, we have discussed collider bias in fairness assessment. We have presented theoretical results based on the graphical causal model to avoid collider

biases in fairness assessment. The results are useful for discrimination detection and also for feature selection for building fair classifiers. We have proposed an Unbiased Situation Test (UST) algorithm for the fairness assessment of a classifier without accessing the training data set or a sample of the population. Experimental results show that UST effectively reduces collider biases and can be used to assess the fairness of a classifier without accessing to data. The UST is promising for an audit agency to audit machine learning models by private companies and organisations.

A Additional Definition and Theorem

Definition 8 (d-separation [17]). *A path π in a DAG is said to be d-separated (or blocked) by a set of nodes $\mathbf{Z}$ iff (1) π contains a chain $V_i \to V_k \to V_j$ and a fork $V_i \leftarrow V_k \to V_j$ node such that the middle node V_k is in $\mathbf{Z}$, or (2) π contains a collider V_k such that V_k is not in $\mathbf{Z}$ and no descendant of V_k is in $\mathbf{Z}$.*

Theorem 2 (Rules of do-Calculus [17]). *Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$ be arbitrary disjoint sets of variables in a causal DAG $\mathcal{G}$. The following rules hold, where $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}$ are the shorthands of $\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}$ and $\mathbf{W} = \mathbf{w}$ respectively.*

Rule 1. (Insertion/deletion of observations):
$P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{w})$, *if* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})$ *in* $\mathcal{G}_{\overline{\mathbf{X}}}$.
Rule 2. (Action/observation exchange):
$P(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, \mathbf{w})$, *if* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})$ *in* $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$.
Rule 3. (Insertion/deletion of actions):
$P(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{w})$, *if* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})$ *in* $\mathcal{G}_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}$, *where $\mathbf{Z}(\mathbf{W})$ is the nodes in $\mathbf{Z}$ that are not ancestors of any node in $\mathbf{W}$ in* $\mathcal{G}_{\overline{\mathbf{X}}}$.

B Proofs

B.1 Proof of Theorem 1

Theorem 1. DAG $\mathcal{G}$ contains variables A and Y, and variable set $\mathbf{X}$ where $(A \cup Y) \cap \mathbf{X} = \emptyset$. The causal sufficiency is satisfied. $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Pa'(Y) = \mathbf{pa})$ where $Pa'(Y)$ is the set of all parents of Y in $\mathcal{G}$ excluding A.

Proof. Firstly, let $\mathbf{X} = \{\mathbf{C} \cup \mathbf{Q}\}$ where $\mathbf{C}$ contains descendant nodes of Y, and $\mathbf{Q}$ contains non-descent nodes of Y. We have $P(y \mid do(A = a), do(\mathbf{C} = \mathbf{c}), do(\mathbf{Q} = \mathbf{q})) = P(y \mid do(A = a), do(\mathbf{Q} = \mathbf{q}))$. This is achieved by repeatedly using Rule 3 of Theorem 2. We show this by an example where $C \in \mathbf{C}$, $P(y \mid do(A = a), do(C = c), do(\mathbf{Q} = \mathbf{q})) = P(y \mid do(A = a), do(\mathbf{Q} = \mathbf{q}))$ because $Y \perp\!\!\!\perp C$ in DAG $\mathcal{G}_{\overline{A},\overline{C}}$ where the incoming edges to A and to C have been removed.

Secondly, we consider $P(y \mid do(A = a), do(\mathbf{Q} = \mathbf{q}))$ only. Based on the Markov condition 1, Y is independent of all its non-descendant nodes given

its parents. Therefore, $P(y \mid do(A = a), do(\mathbf{Q} = \mathbf{q})) = P(y \mid do(A = a), do(Pa'(Y) = \mathbf{pa}))$.

Thirdly, we will prove $P(y|do(A = a), do(Pa'(Y) = \mathbf{pa})) = P(y \mid A = a, Pa'(Y) = \mathbf{pa})$. This can be achieved by repeatedly applying Rule 2 of Theorem 2.

Let $Pa(Y) = \{A, X_1, X_2, \ldots, X_k\}$.

$$
\begin{aligned}
&P(y \mid do(A = a), do(X_1 = x_1), do(X_2 = x_2), \ldots, do(X_k = x_k)) \\
&= P(y \mid A = a, do(X_1 = x_1) do(X_2 = x_2), \ldots, do(X_k = x_k)) \\
&\textit{Since } Y \perp\!\!\!\perp A | X_1, X_2, \ldots, X_k \textit{ in } \mathcal{G}_{\overline{X_1}, \overline{X_2}, \ldots, \overline{X_k}, \underline{A}} \\
&= P(y \mid A = a, X_1 = x_1, do(X_2 = x_2), \ldots, do(X_k = x_k)) \\
&\textit{Since } Y \perp\!\!\!\perp X_1 | A, X_2, \ldots, X_k) \textit{ in } \mathcal{G}_{\overline{X_2}, \ldots, \overline{X_k}, \underline{X_1}} \\
&\textit{Repeat } k - 1 \textit{ times} \\
&= P(y \mid A = a, X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) \\
&= P(y \mid A = a, Pa'(Y) = \mathbf{pa})
\end{aligned}
$$

Now, we get,

$$P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Pa'(Y) = \mathbf{pa})$$

B.2 Proof of Corollary 1

Corollary 1. Let $Ca(Y)$ include all the direct causes and only direct causes of Y except A. $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Ca(Y) = \mathbf{ca})$.

Proof. Direct causes of Y will be parent nodes of Y in any DAG even when the unobserved common causes are included, i.e. the causal sufficiency is unsatisfied. Since $Pa'(Y) = Ca(Y)$ and there is not an unobserved variable in between a direct cause and Y, $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Ca(Y) = \mathbf{ca})$ can be derived following the same procedure in Theorem 1.

Since other variables apart from $Pa'(Y)$ are not used in reducing $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x}))$, the unobserved common casues between these variables are irrelevant to the deduction and do not affect the above conclusion.

B.3 Proof of Corollary 2

Corollary 2. Let $\mathbf{B}$ include all the direct causes and some (or all) indirect causes of Y. We have $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, \mathbf{B} = \mathbf{b})$.

Proof. Let $\mathbf{B} = Ca'(Y) \cup \mathbf{R}$, and $\mathbf{R}$ includes indirect causes of Y. Following Corollary 1, $P(y \mid do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y \mid A = a, Ca'(Y) = \mathbf{ca})$. Based on the Markov condition 1, Y is independent of $\mathbf{R}$ given $A \cup Ca'(Y)$. Hence, $P(y \mid A = a, Ca'(Y) = \mathbf{ca}) = P(y \mid A = a, \mathbf{B})$.

References

1. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007)
2. Bendick, M.: Situation testing for employment discrimination in the United States of America. Horizons stratégiques **3**, 17–39 (2007)
3. Caton, S., Haas, C.: Fairness in machine learning: a survey (2020). arXiv preprint arXiv:2010.04053
4. Chiappa, S.: Path-specific counterfactual fairness. In: AAAI, pp. 7801–7808 (2019)
5. Cole, S.R., et al.: Illustrating BIAS due to conditioning on a collider. Int. J. Epidemiol. **39**(2), 417–20 (2010)
6. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning (2018). arXiv preprint arXiv:1808.00023
7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
8. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: KDD, pp. 259–268 (2015)
9. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: NeurIPS, pp. 3315–3323 (2016)
10. Hernán, M.A., Robins, J.M.: Causal Inference: What If. Chapman & Hall/CRC, Boca Raton (2020)
11. Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: NeurIPS, pp. 656–666 (2017)
12. Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: NeurIPS, pp. 4069–4079 (2017)
13. Li, J., Liu, J., Liu, L., Le, T.D., Ma, S., Han, Y.: Discrimination detection by causal effect estimation. In: BigData, pp. 1087–1094. IEEE (2017)
14. Lohia, P.K., Ramamurthy, K.N., Bhide, M., Saha, D., Varshney, K.R., Puri, R.: Bias mitigation post-processing for individual and group fairness. In: ICASSP, pp. 2847–2851. IEEE (2019)
15. Luong, B.T., Ruggieri, S., Turini, F.: K-NN as an implementation of situation testing for discrimination discovery and prevention. In: KDD, pp. 502–510 (2011)
16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on BIAS and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
17. Pearl, J.: Causality. Cambridge University Press (2009)
18. Pearl, J., Mackenzie, D.: The Book of Why. Basic Books, New York (2018)
19. Speicher, T., et al.: A unified approach to quantifying algorithmic unfairness: measuring individual & group unfairness via inequality indices. In: KDD, pp. 2239–2248 (2018)
20. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, Prediction, And Search. MIT Press (2000)
21. Wu, Y., Zhang, L., Wu, X., Tong, H.: Pc-fairness: a unified framework for measuring causality-based fairness. In: NeurIPS, vol. 32 (2019)
22. Xu, Z., et al.: Assessing Classifier Fairness With Collider Bias (2022). arXiv preprint arXiv:2010.03933
23. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW, pp. 1171–1180 (2017)

24. Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. **4**(1), 1–16 (2017). https://doi.org/10.1007/s41060-017-0058-x
25. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI, pp. 2718–2724 (2016)
26. Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI, pp. 3929–3935 (2017)